Rapport de Projet – IA

Projet A3 : Accidentologie

*Développement d’une application d’étude des incidents de la route en 2009*

*Tristan Saëz – Adrien Leboucher – Vincent Le Brenn*

*Année 2022 - 2023*

Table des matières

[Préambule 3](#_Toc136513218)

[Gestion de projet & Environnement de travail 3](#_Toc136513219)

[Découverte & Préparation des données 4](#_Toc136513220)

[Découverte des données 4](#_Toc136513221)

[Préparation des données 5](#_Toc136513222)

[Apprentissage non-supervisé 6](#_Toc136513223)

[Réduction de dimension 6](#_Toc136513224)

[Clustering 7](#_Toc136513225)

Apprentissage supervisé

Répartition des données

Classification kNN

Classification "haut niveau"

Evaluation des résultats

Quantitative Supervisé

Quantitative Non-Supervisé

# Préambule

Ce rapport est le résultat des travaux effectués lors de notre projet en Intelligence Artificielle effectué du 30 mai 2023 au 2 juin 2023. Il consiste en l’étude des accidents de la route en 2009 et au développement d’un projet d’apprentissage automatique en lien avec ces accidents.

Le projet se divise en plusieurs partie : Préparation et Visualisation des données, Apprentissage non-supervisé (clustering), Apprentissage supervisé, Evaluation et Analyse des résultats et Création de scripts d’automatisation via commandes bash.

Les objectifs de ce projet furent les suivants :

• Comprendre les étapes d’un projet d’apprentissage automatique

• Apprendre à évaluer une méthode d’apprentissage automatique

• Manipuler les différentes techniques d’apprentissage automatique :

• Apprentissage non supervisé : réduction de dimensionnalité, clustering

• Apprentissage supervisé : régression, classification

La prochaine partie est dédiée à l’environnement de travail et la gestion de projet

# Gestion de projet & Environnement de travail

La première partie fut dédiée à la configuration de notre environnement de travail :

* Un dépôt GIT utilisé pour la programmation collaborative
* Un diagramme de Gantt pour le partage des tâches

Dépôt Git : [https:/github.com/TrLSeysh/ProjetCIR3\_IA](https://github.com/TrLSeysh/ProjetCIR3_IA)

# Découverte & Préparation des données

Comme n’importe quelle base de données, elle ne fut pas exploitable sans la préparer. Plusieurs étapes furent nécessaires lors de la préparation du jeu de données pour le rendre exploitable :

Découverte

* Comprendre la base de données et son organisation
* Définir les valeurs utiles et subsidiaires

Préparation

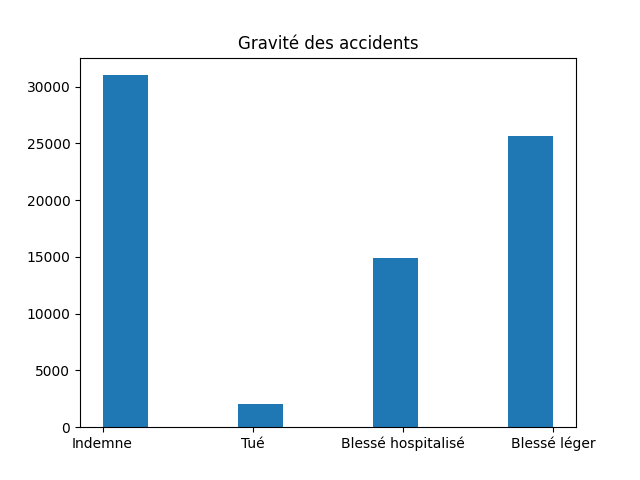
* Convertir les valeurs non-numériques
* Mettre les dates et heures au bon format

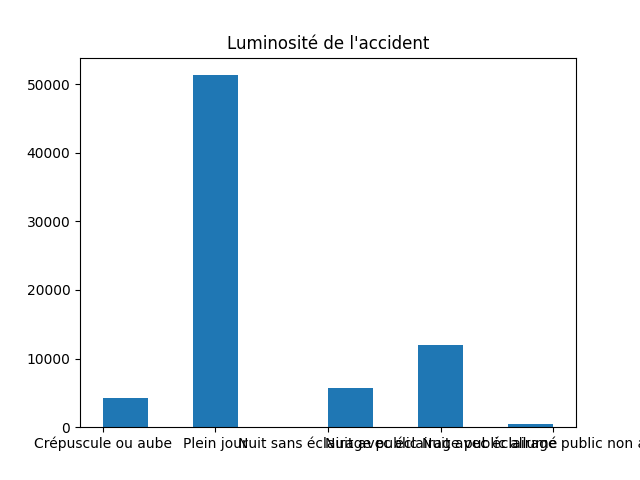
## Découverte des données

Nous avons tout d’abord étudié l’élément : {descr\_grav} qui représente la gravité de l’accident. Nous en avons déduit qu’un regroupement de label pouvait être judicieux pour n’en garder que 3 :

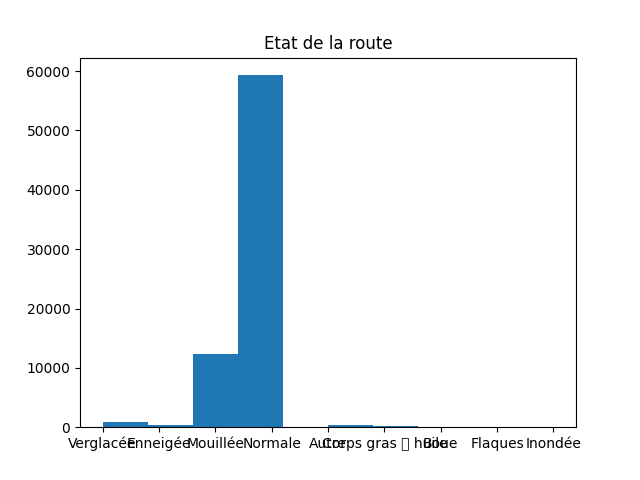
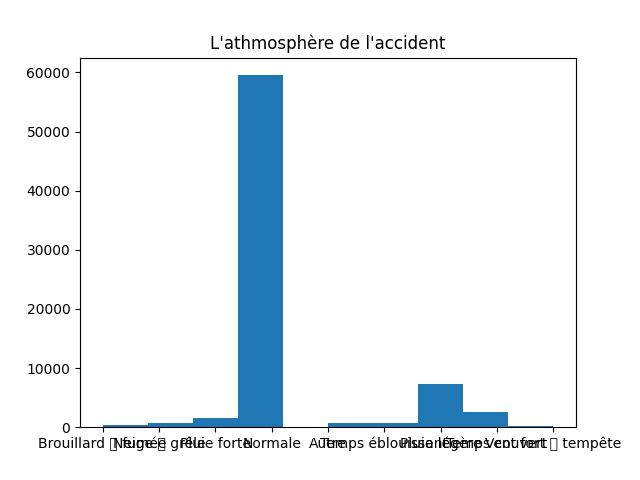
* Indemne
* Blessé léger
* Accident grave (Tué & Blessé Hospitalisé)

Ceci permettra de combler les écarts entre les labels pour avoir une précision accrue.

Ensuite, nous avons étudié le paramètre {descr\_lum} qui décrit la luminosité lors de l’accident. Bien que ce trait puisse être intéressant à étudier, nous avons décidé de ne pas le modifier pour se concentrer sur l’élément le plus important de la base : la gravité de l’accident.



Notre objectif est avant tout de prédire la gravité de l’accident, de ce fait tous les autres traits, bien que pertinent pour préciser la prédiction des accidents, ils ont été mis de côté. Le constat est donc le même pour les paramètres {descr\_athmo} et {descr\_etat\_surf} qui ont, eux, été retirés de la base d’apprentissage.



## Préparation des données

2 modifications principales ont dû être effectuée sur la base de données pour permettre une exploitation idéale de son contenu. La première fut la conversion des valeurs non-numériques en entiers.

Pour se faire, il a d’abord fallu récupérer chaque colonne et vérifier le type de chacune d’elles. Nous avons utilisé la fonction isinstance() de python qui permet de vérifier le type d’une variable. Il était également nécessaire de ne pas modifier certaines colonnes telles que le code\_insee et la date qui sont traités différemment.

Ensuite, la fonction factorize() de la library pandas nous a permis de donner à chaque valeur non-numérique un entier.

Enfin, pour garder en mémoire la valeur liée à chaque indice nous les avons intégré à un fichier excel via la fonction to\_excel(). Le résultat est le suivant :

Une image contenant texte, nombre, capture d’écran, Parallèle

Description générée automatiquement

Premières lignes du fichier de conversion obtenu 1

La seconde modification nécessaire fut le changement de format des dates. Ici, nous avons dû utiliser la library datetime et ses fonctions .month et .hour qui permettent de récupérer le mois et l’heure de l’accident (l’année étant toujours 2009, il n’était pas intéressant de la conserver) et de les mettre au bon format datetime.

Nous avons donc ajouté 2 colonnes mois et heure puis supprimé la colonne date.

# Apprentissage non-supervisé

La première partie de notre réseau d’apprentissage automatique fut l’apprentissage non-supervisé et plus précisément les méthodes de clustering des données. Mais pour pouvoir faire des clusters cohérents de manière efficace, il y eut la nécessité de réduire le nombre de features et de corréler certains attributs entre eux : réduire les dimensions de la base de données.

## Réduction de dimension

Nous avons commencé par calculer le coefficient de corrélation en fonction de la gravité de l’accident (feature maîtresse de l’apprentissage). Voici les résultats :

|  |  |  |
| --- | --- | --- |
| Coefficient de corrélation en fonction de la gravité des accidents | | |
| feature | Coef de corrélation | Supprimée ? |
| jour | 0.01671 | ❌ |
| id\_usa | 0.001948 | ❌ |
| latitude | 0.006072 |  |
| longitude | 0.008392 |  |
| descr\_athmo | 0.012155 | ❌ |
| id\_code\_insee | 0.012865 | ❌ |
| descr\_etat\_surf | 0.015064 | ❌ |
| descr\_agglo | 0.016364 | ❌ |
| description\_intersection | 0.019234 |  |
| heure | 0.025678 | ❌ |
| descr\_lum | 0.030840 |  |
| num\_veh | 0.034247 | ❌ |
| descr\_motif\_traj | 0.054042 | ❌ |
| descr\_type\_col | 0.055053 |  |
| X (id) | 0.061824 | ❌ |
| place | 0.110989 | ❌ |
| an\_nais | 0.132073 |  |
| descr\_dispo\_secu | 0.222666 | ❌ |
| descr\_cat\_veh | 0.239771 |  |
| descr\_grav | 1.000000 |  |

Le coefficient de corrélation est contenu entre 1 et -1. Plus la valeur s’approche d’un extrême, plus la corrélation est forte où, proche de 1, elle est très corrélée positivement, et proche de -1, elle est très corrélée négativement. Plus le coefficient se rapproche de 0 moins la corrélation est forte et moins les variables s'influencent les unes aux autres. Ici on peut voir qu’à partir de X(id) la corrélation est faible. Mais encore plus faible à partir de longitude.

La colonne « Supprimée ? » montre les features qui ont été supprimées par la suite

Pour représenter les différentes corrélations nous avons décidé de les mettre sous forme de matrices, ce qui a donné les résultats suivant en comparant avant et après réduction de dimension :

Une image contenant diagramme, capture d’écran, Caractère coloré, carré

Description générée automatiquementUne image contenant capture d’écran, diagramme, texte, Caractère coloré

Description générée automatiquement

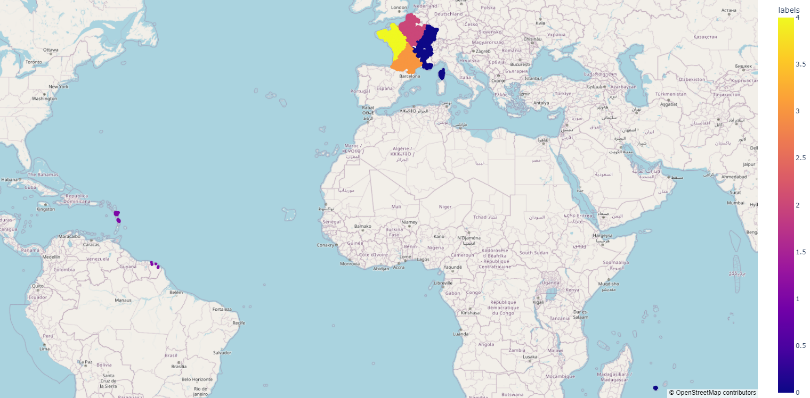
Matrice de corrélation avant réduction de dimension

Matrice de corrélation après réduction de dimension

## Clustering

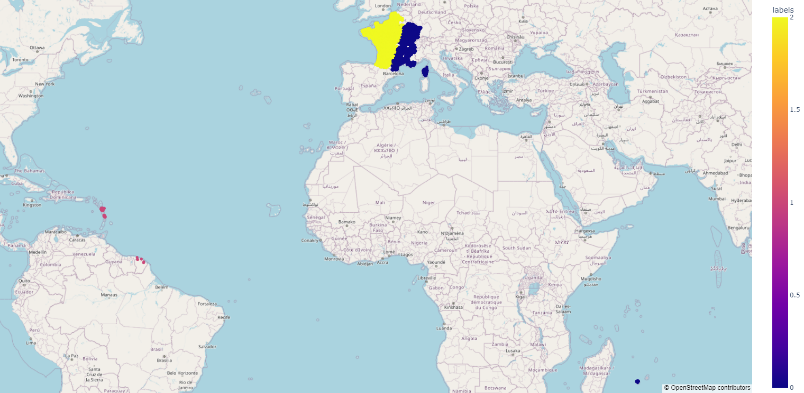
Nous avons décomposé le clustering en 2 étapes, une première avec un calcul dit « from scratch » et une seconde en utilisant la bibliothèque scikit-learn pour comparer les résultats obtenus.

Pour obtenir des résultats pertinents, nous avons fait varier le nombre de cluster : n = [3, 5, 8, 12] et également la méthode de calcul « from scratch » : méthode = [L1, L2, Haversine].

Voici les résultats obtenus lors du clustering avec les différentes méthodes pour 8 clusters :

*Méthode : Haversine*

*Méthode : Scikit-learn*



*Méthode : Haversine*

*Méthode : L1*

*Méthode : L2*

Après avoir obtenu des résultats sur un nombre de clusters variable nous les avons analysé avec des indicateurs de rendement :

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nombre de cluster | | | | | | | | | | | |
| 3 | | | | | | | | | | | |
| Méthode | | | scikit | | L1 | | L2 | | Haversine | | |
| Metric | | | Score | | Score | | Score | | Score | | |
| Coefficient de silhouette | | | 0.4396002394042591 | | 0.3099044630729121 | | 0.3066844697028553 | | 0.3066844697028553 | | |
| Index de Calinski Harabasz | | | 1002.1034962689257 | | 619.8313597309689 | | 617.8453619262061 | | 617.8453619262061 | | |
| Index de Bouldin Davies | | | 0.759965228551648 | | 1.4849925933563453 | | 1.4822945609067348 | | 1.4822945609067348 | | |
|  |  |  |  |  |  |  |  |  |  |  |
| Nombre de cluster | | | | | | | | | | | |
| 5 | | | | | | | | | | | |
| Méthode | | | scikit | | L1 | | L2 | | Haversine | | |
| Metric | | | Score | | Score | | Score | | Score | | |
| Coefficient de silhouette | | | 0.44375880274301455 | | 0.20855898682664645 | | 0.3088980353693082 | | 0.3088980353693082 | | |
| Index de Calinski Harabasz | | | 1242.0986175669645 | | 407.9566647218841 | | 790.2322194733962 | | 790.2322194733962 | | |
| Index de Bouldin Davies | | | 0.7179728279735788 | | 1.9968265749287402 | | 0.8234573037440249 | | 0.8234573037440249 | | |
|  |  |  |  |  |  |  |  |  |  |  |
| Nombre de cluster | | | | | | | | | | | |
| 8 | | | | | | | | | | | |
| Méthode | | | scikit | | L1 | | L2 | | Haversine | | |
| Metric | | | Score | | Score | | Score | | Score | | |
| Coefficient de silhouette | | | 0.41307080643298366 | | 0.26924551509454425 | | 0.29935651624751863 | | 0.29935651624751863 | | |
| Index de Calinski Harabasz | | | 1230.1642992694271 | | 642.0362689580388 | | 700.0842140392815 | | 700.0842140392815 | | |
| Index de Bouldin Davies | | | 0.8160015678900034 | | 1.033938304118349 | | 0.9646450498949747 | | 0.9646450498949747 | | |
|  |  |  |  |  |  |  |  |  |  |  |
| Nombre de cluster | | | | | | | | | | | |
| 12 | | | | | | | | | | | |
| Méthode | | | scikit | | L1 | | L2 | | Haversine | | |
| Metric | | | Score | | Score | | Score | | Score | | |
| Coefficient de silhouette | | | 0.4622257479910289 | | 0.35067055901472927 | | 0.28521006951152567 | | 0.28521006951152567 | | |
| Index de Calinski Harabasz | | | 1366.3215155872706 | | 851.2028692835758 | | 724.7676272093031 | | 724.7676272093031 | | |
| Index de Bouldin Davies | | | 0.7528093020232839 | | 0.8516612387507877 | | 1.2022495277774983 | | 1.2022495277774983 | | |

